Non-coding RNAs: how to find and make sense of them

Judith Zaugg

Luscombe group at EBI

zaugg@ebi.ac.uk

Outline of this lecture

- Introduction about non-coding RNAs
- Methods to identify non-coding RNAs
- Methods to assign biological/molecular functions to non-coding RNAs

INTRODUCTION

“Although less than 2% of a mammalian genome codes for protein, studies consistently show that half or even more of the genome is transcribed”

Long noncoding RNAs: the search for function

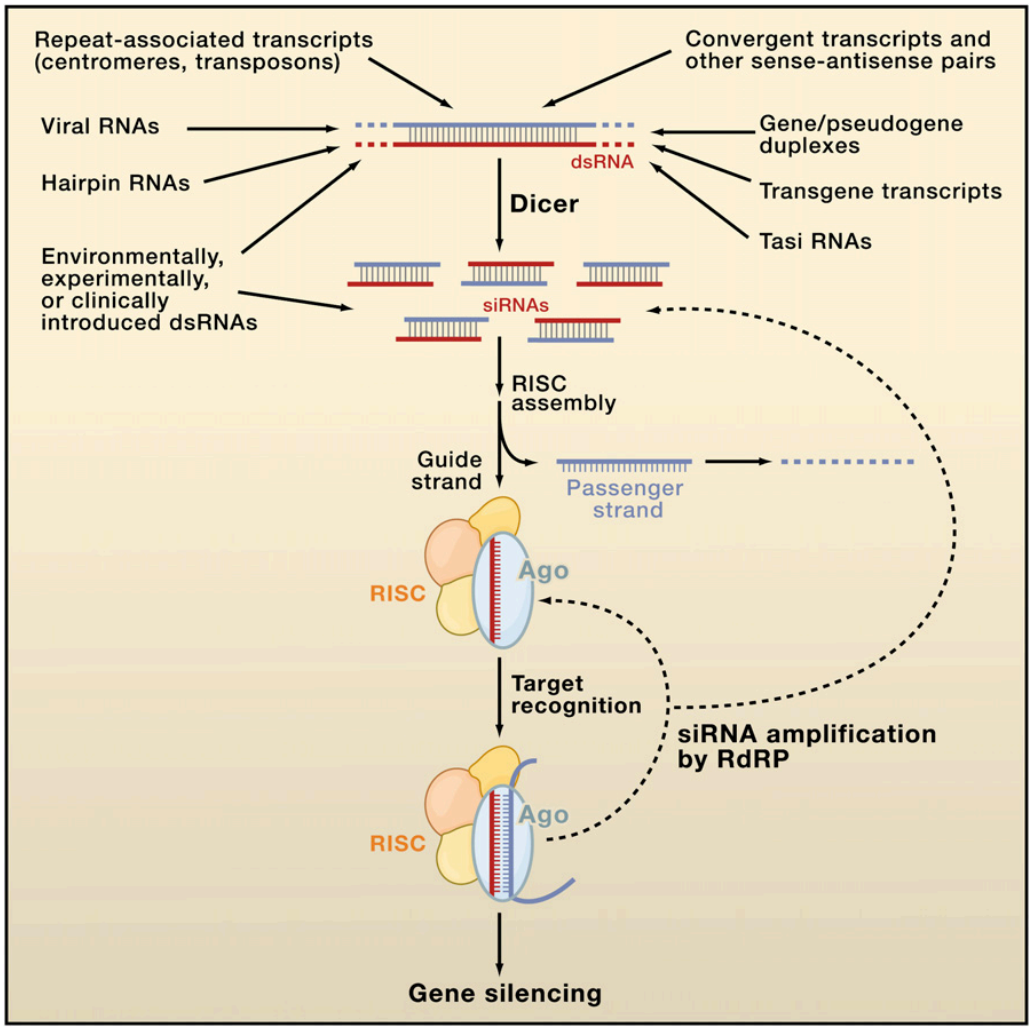
Monya Baker

What are non-coding RNAs

- RNAs that do not encode proteins
- “Textbook knowledge” non-coding (nc)RNA
  - *transfer* (t)RNAs: transfer amino acids
  - *ribosomal* (r)RNAs: constitute the catalytic center of the ribosome
  - *small nucleolar* (sn)RNAs: involved in nuclear processes such as splicing or rRNA processing
- small regulatory RNAs like *small interfering* (si)RNAs
  - gene silencing through double-stranded RNA
- long non-coding RNAs: focus of this presentation

small regulatory RNAs

- just a short overview, not discussed here in detail



Cell

Leading Edge
Review

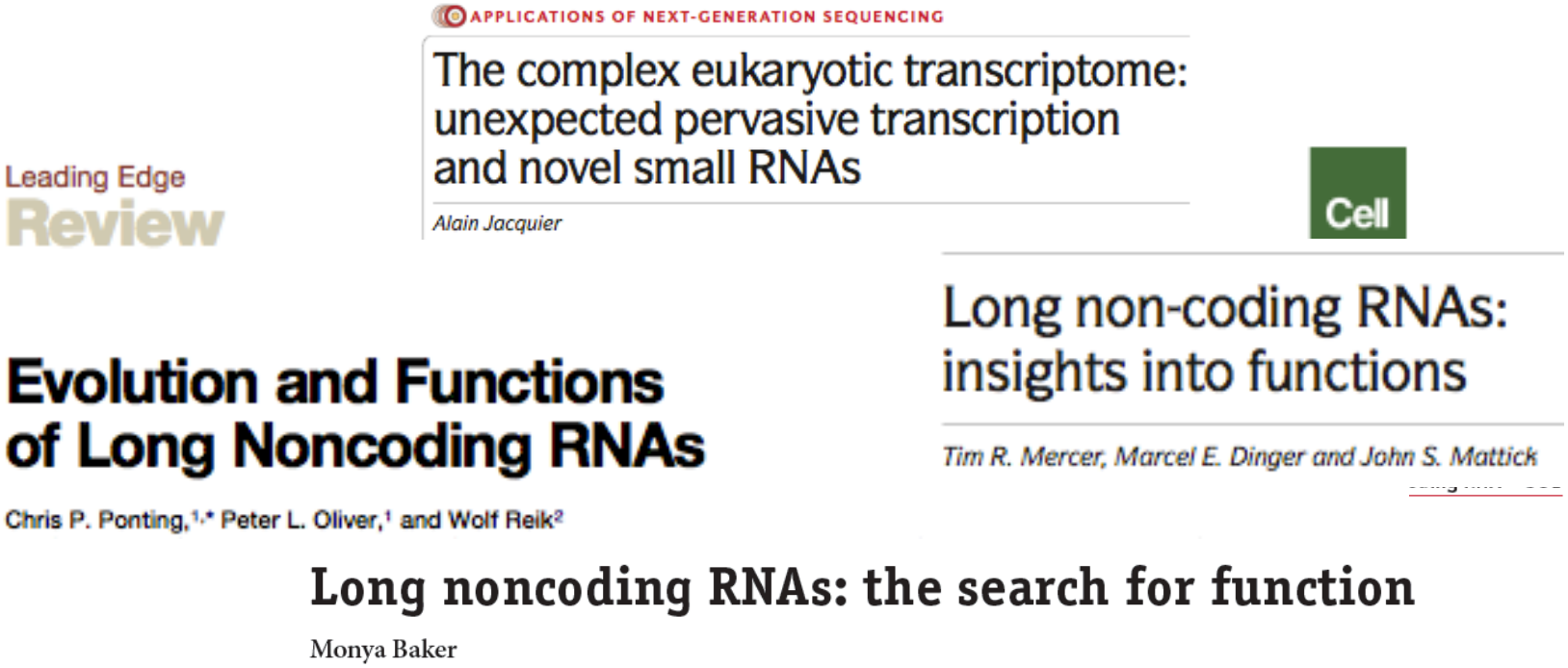
Origins and Mechanisms of miRNAs and siRNAs

Richard W. Carthew[1,*] and Erik J. Sontheimer[1,*]
[1]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2205 Tech Drive, Evanston, IL 60208-3500, USA
*Correspondence: r-carthew@northwestern.edu (R.W.C.), erik@northwestern.edu (E.J.S.)
DOI 10.1016/j.cell.2009.01.035

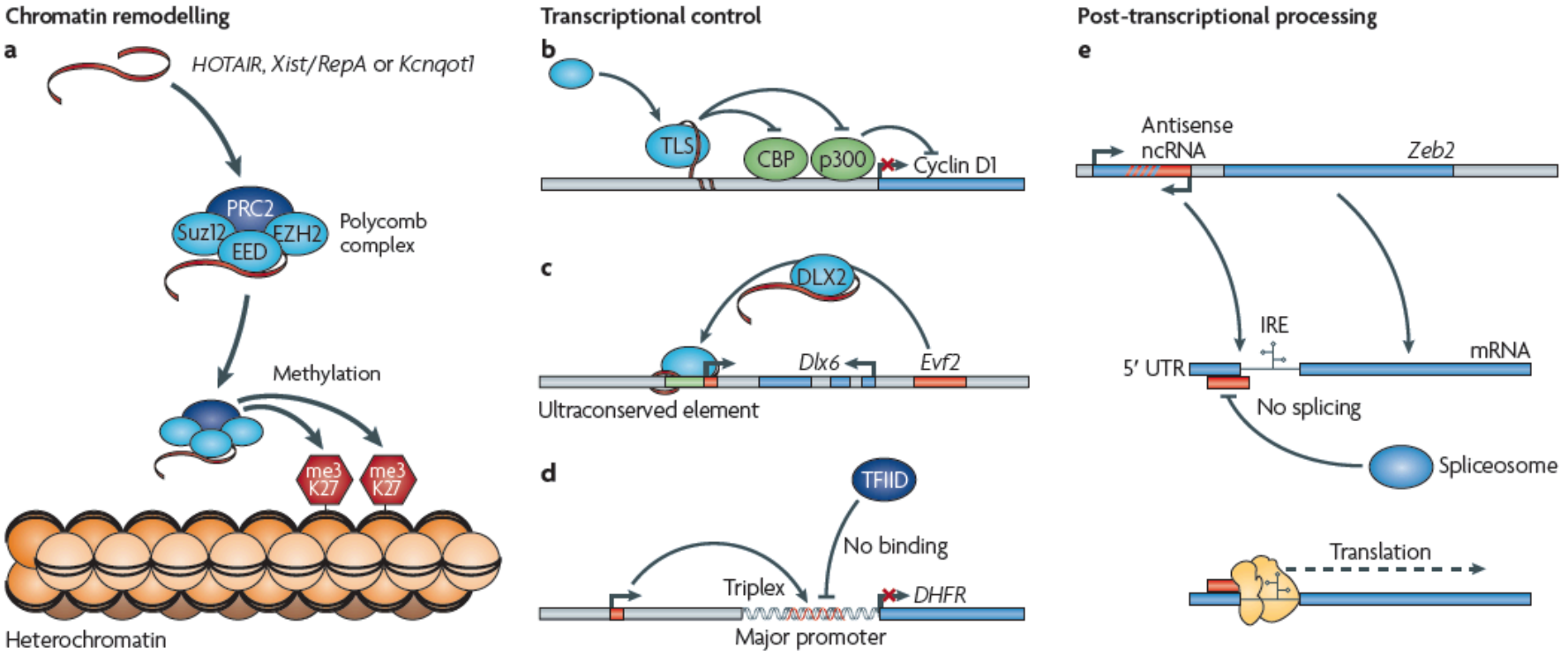
long non-coding RNAs (>200bp)

Recent reviews of long non-coding RNAs have focused on:

- function of ncRNAs and mainly describe examples
- how ncRNAs are transcribed



Examples of how ncRNAs can function

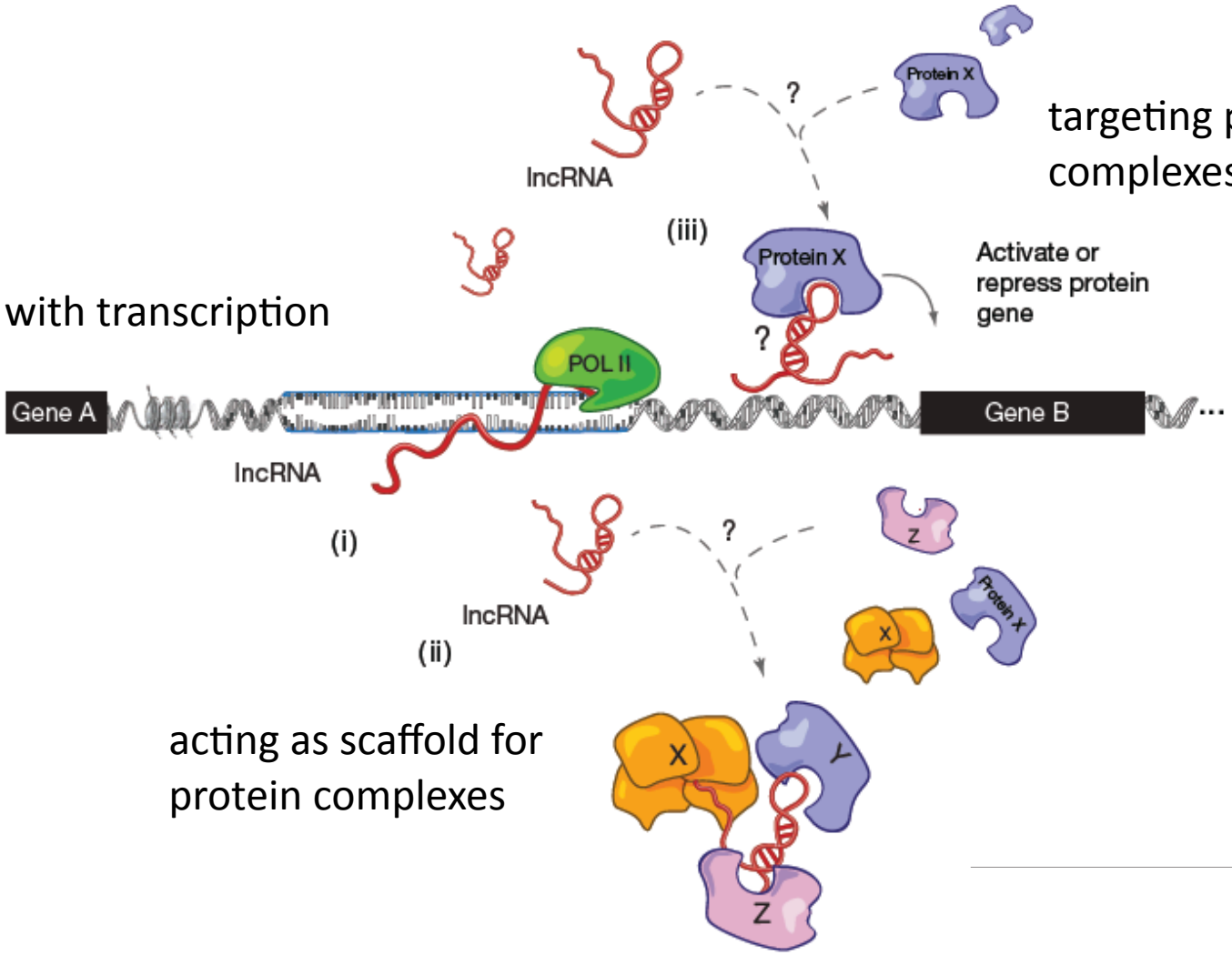


Long non-coding RNAs: insights into functions

Tim R. Mercer, Marcel E. Dinger and John S. Mattick

Molecular mechanisms of how ncRNAs can function

Interference with transcription



targeting protein complexes to DNA

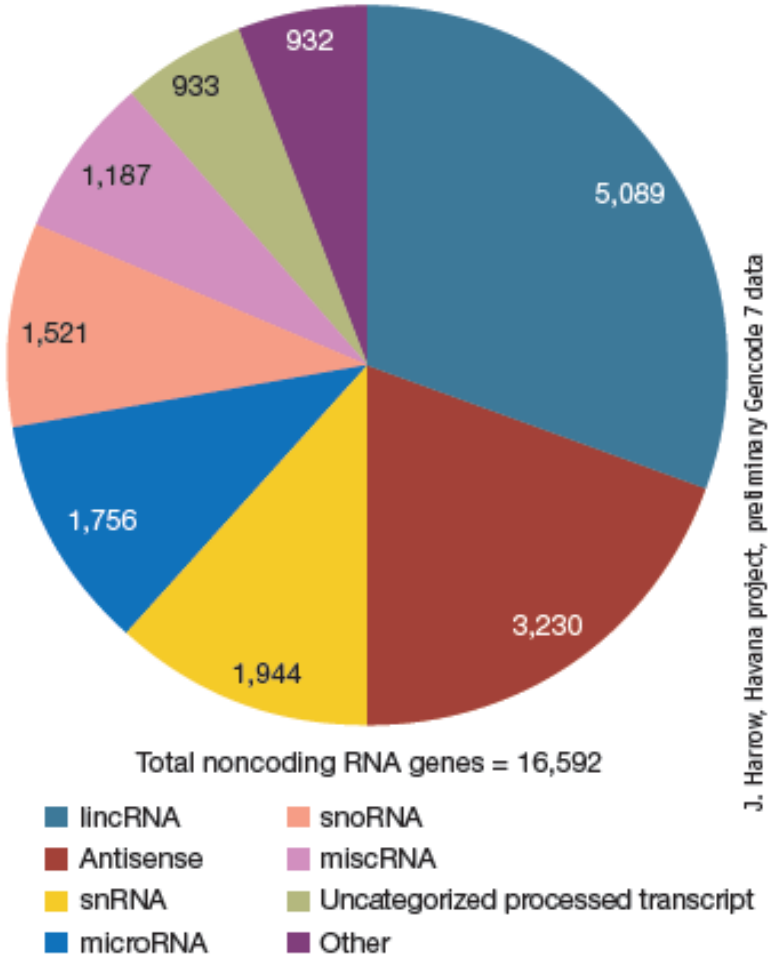
acting as scaffold for protein complexes

TECHNOLOGY FEATURE

Long noncoding RNAs: the search for function

Monya Baker

Distribution of ncRNAs



long ncRNAs are similar to mRNAs:

- RNA polymerase II-promoted
- polyadenylated
- often alternatively spliced

TECHNOLOGY FEATURE

Long noncoding RNAs: the search for function

Monya Baker

FOCUS OF THIS REVIEW

Focus of this review

Many studies have identified novel ncRNAs, tried to classify them, and assign biological functions to them. Various approaches have been developed these steps.

Here I will review the different approaches to

1. identify ncRNAs
2. assign a biological function or molecular mechanism to ncRNAs

To get a better idea of how different studies can be compared

Very general overview of approaches to identify ncRNAs

BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 8. NO 6. 437–443 doi:10.1093/bfgp/elp041

Annotating non-coding transcription using functional genomics strategies

Alistair R. R. Forrest, Rehab F. Abdelhamid and Piero Carninci

Abstract

Non-coding RNA (ncRNA) transcripts are RNA molecules that do not code for proteins, but elicit function by other mechanisms. The vast majority of RNA produced in a cell is non-coding ribosomal RNA, produced from relatively few loci, however more recently complementary DNA (cDNA) cloning, tag sequencing, and genome tiling array studies suggest that ncRNAs also account for the majority of RNA species produced by a cell. ncRNA based regulation has been referred to as a 'hidden layer' of signals or 'dark matter' that control gene expression in cellular processes by poorly described mechanisms. These terms have appeared as ncRNAs until recently have been ignored by expression profiling and cDNA annotation projects and their mode of action is diverse (e.g. influencing chromatin structure and epigenetics, translational silencing, transcriptional silencing). Here, we highlight recent functional genomics strategies toward identifying and assigning function to ncRNA transcription.

Keywords: *non-coding RNA; Sequencing; transcription; annotation*

IDENTIFICATION OF NON-CODING RNA

1. Identification and validation of functional ncRNA

a) Identification
   I. Identification based on template-free expression measurement
   II. Identification based on chromatin-state signatures
   III. Identification of unstable transcripts
b) Validation and how to distinguish coding from non-coding
   I. conservation
   II. distinguish coding from non-coding RNA
c) Application: Computational identification based on published data

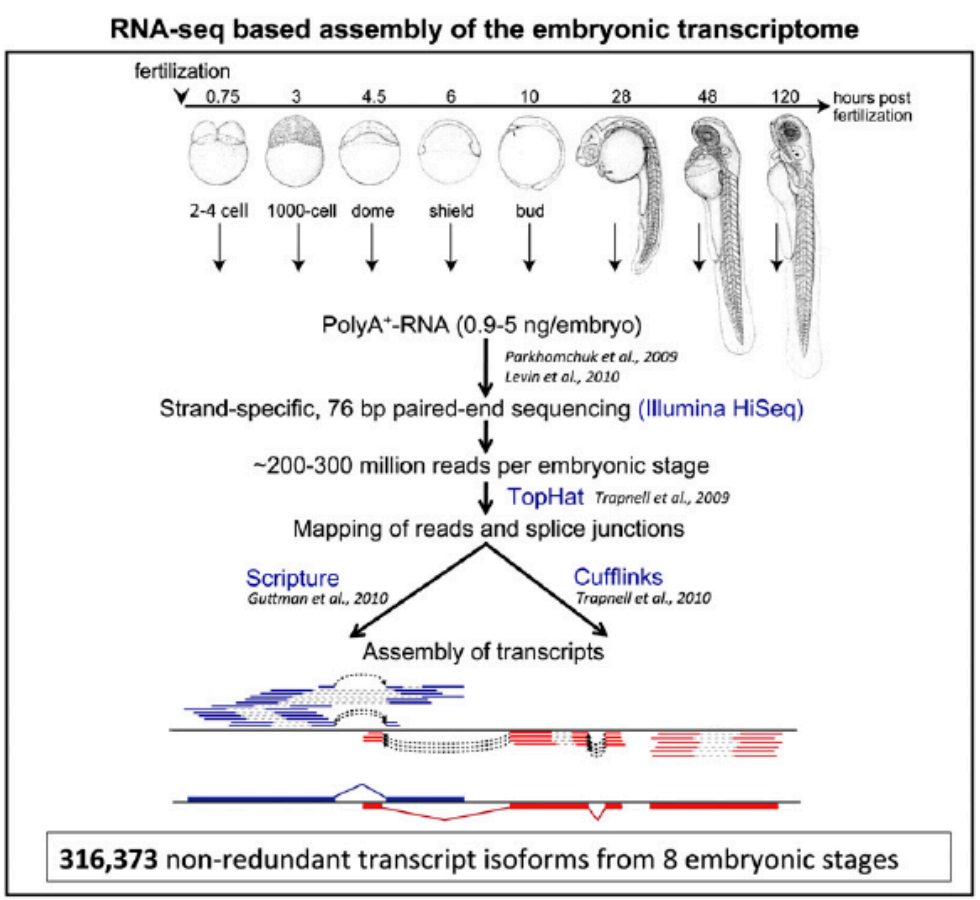
- RNA-seq
- tiling arrays
- global run-on (GRO)-seq: detection of nascent transcripts

a) I. RNA-seq: zebrafish

Resource

Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis

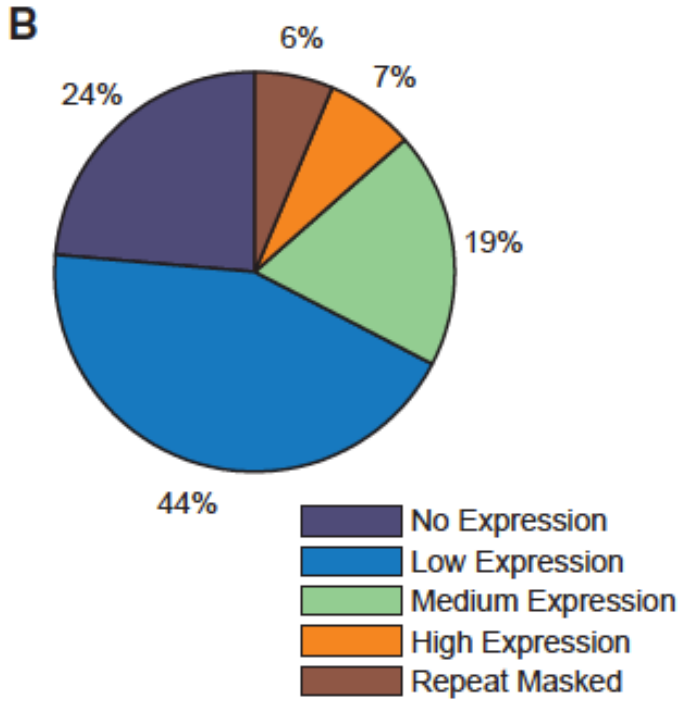
Andrea Pauli,[1,7,8] Eivind Valen,[2,7] Michael F. Lin,[3,4] Manuel C
Nadine L. Vastenhouw,[1] Joshua Z. Levin,[4] Lin Fan,[4] Albin Sar
John L. Rinn,[4,5] Aviv Regev,[3,4,6,8] and Alexander F. Schier[1,4,8]



a) I. RNA-seq: first studies in yeast

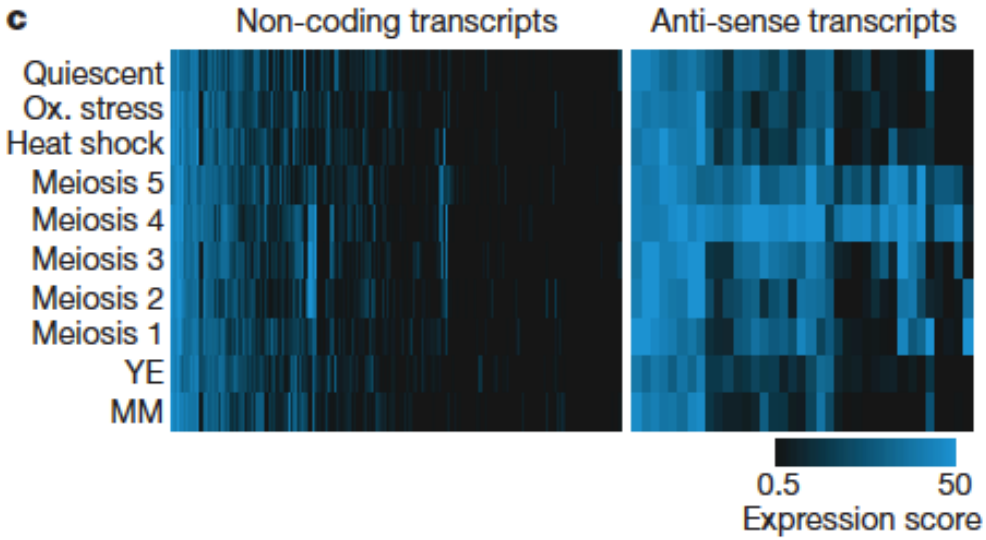
The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1]* Zhong Wang,[1]* Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†



Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution

Brian T. Wilhelm[1]*†, Samuel Marguerat[1]*†, Stephen Watt[1]†, Falk Schubert[1]†, Valerie Wood[1], Ian Goodhead[1]†, Christopher J. Penkett[1]†, Jane Rogers[1] & Jürg Bähler[1]†

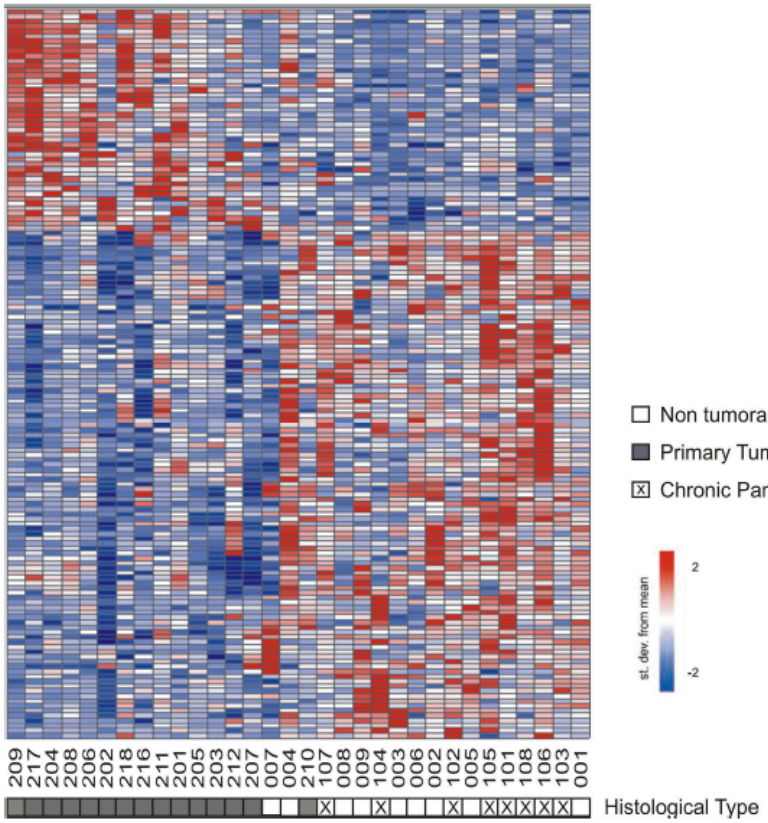


a) I. RNA-seq: humans

RESEARCH Open Access

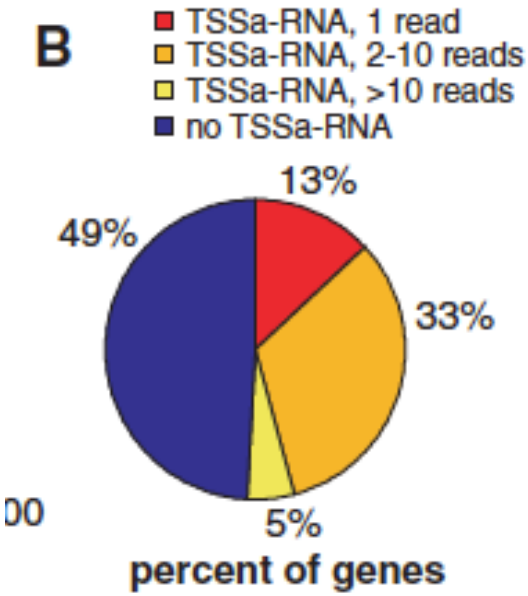
Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer

Ana C Tahira[1], Márcia S Kubrusly[2], Michele F Faria[1], Bianca Dazzani[1], Rogério S Fonseca[1], Vinicius Maracaja-Coutinho[1], Sergio Verjovski-Almeida[1], Marcel CC Machado[2] and Eduardo M Reis[1*]



Divergent Transcription from Active Promoters

Amy C. Seila,[1*] J. Mauro Calabrese,[1,2*†] Stuart S. Levine,[3] Gene W. Yeo,[4‡] Peter B. Rahl,[3] Ryan A. Flynn,[1] Richard A. Young,[2,3] Phillip A. Sharp[1,2§]



RNA-seq for identifying ncRNA

+

- identification of novel transcripts
- in principle reference-independent

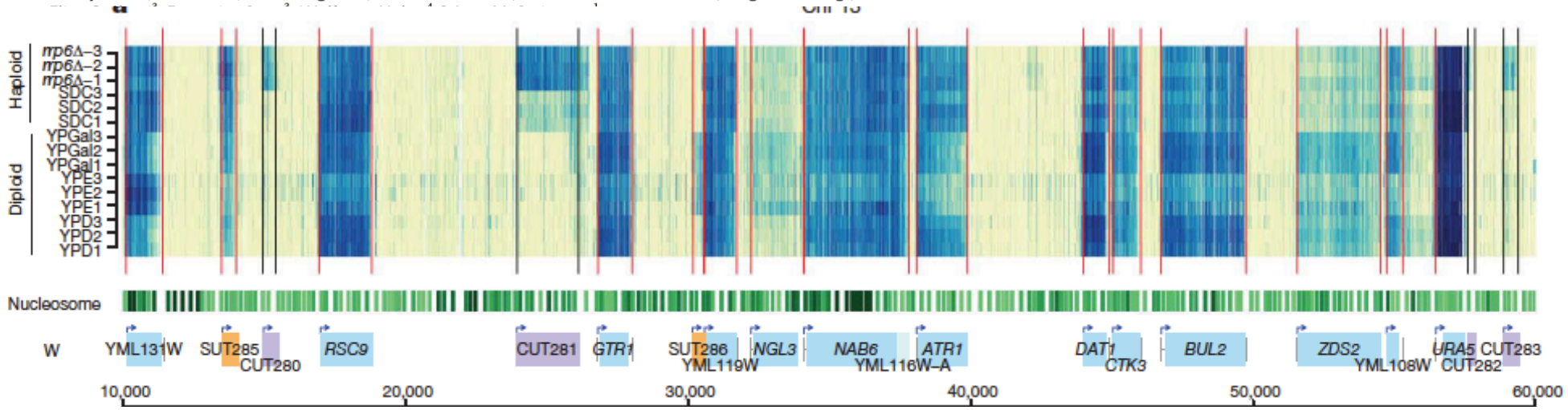
-

- no detection of unstable transcripts
- need high coverage for detecting ncRNAs (they are generally lowly expressed)

a) I. tiling-array

Bidirectional promoters generate pervasive transcription in yeast

Zhenyu Xu[1]*, Wu Wei[1]*, Julien Gagneur[1], Fabiana Perocchi[1], Sandra Clauder-Münster[1], Jurgi Camblong[2],



manual annotation of novel transcripts

tilling-arrays for identifying ncRNA

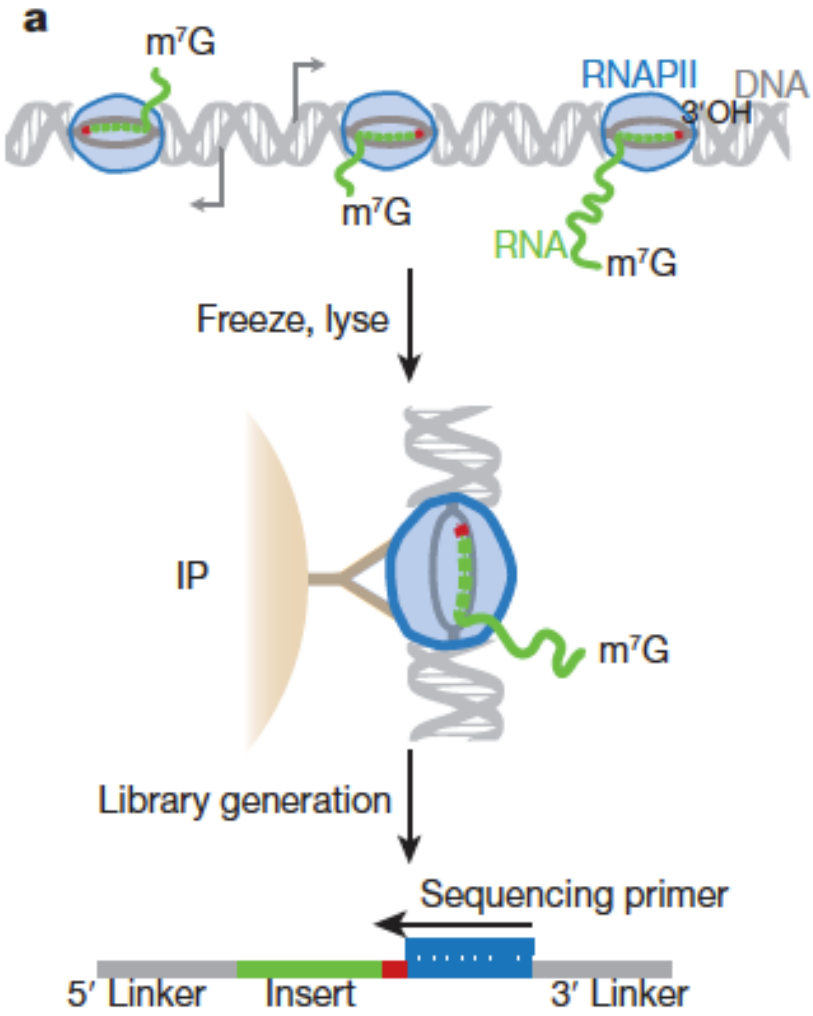
+

- identification of novel transcripts
- for whole genome arrays: reference-independent

-

- no detection of unstable transcripts
- potential difficulty to detect low abundance transcripts (they might fall into the background region)

a) I. GRO-seq (Global run-on)



Nascent transcript sequencing visualizes transcription at nucleotide resolution

L. Stirling Churchman[1] & Jonathan S. Weissman[1]

GRO-seq for identifying ncRNA

+

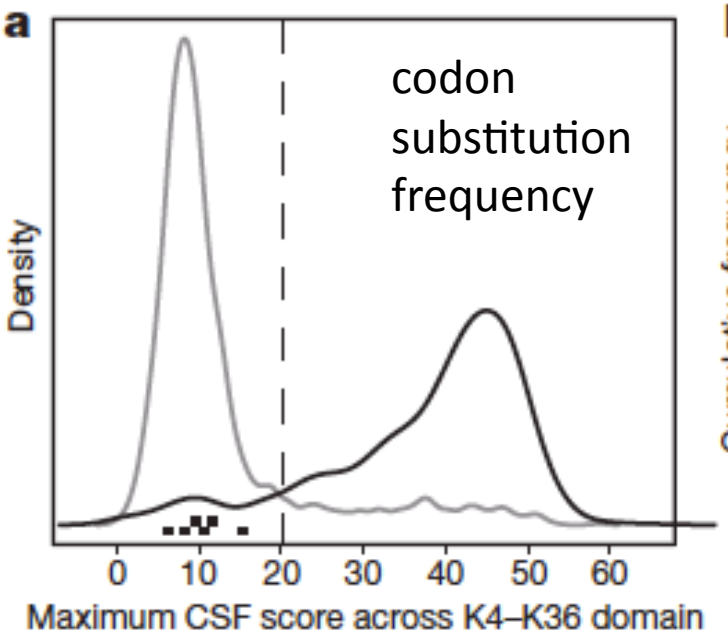
- identification of novel transcripts
- reference-independent
- detection of unstable transcripts

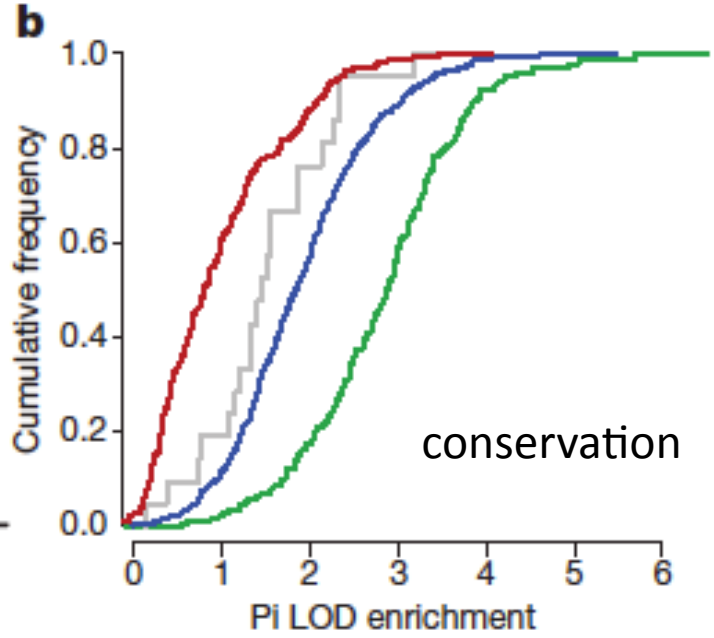
-

- difficult to obtain the transcript lengths
- only unspliced RNAs detected (except for when co-transcriptional splicing occurs)

a) II. Identification based on chromatin state signatures

- active transcription: H3K4me3 at promoter, H3K36me3 across body
- Find K4–K36 regions outside coding genes loci to discover ncRNAs





red: protein coding
blue: lncRNA
green: random regions

Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals

Mitchell Guttman[1,2], Ido Amit[1], Manuel Garber[1], Courtney French[1], Michael F. Lin[1], David Feldser[3], Maite Huarte[1,6], Or Zuk[1], Bryce W. Carey[2,8], John P. Cassady[2,8], Moran N. Cabili[7], Rudolf Jaenisch[2,8], Tarjei S. Mikkelsen[1,4], Tyler Jacks[2,3], Nir Hacohen[1,9], Bradley E. Bernstein[1,10,11], Manolis Kellis[1,5], Aviv Regev[1,2], John L. Rinn[1,6,11]* & Eric S. Lander[1,2,7,8]*

Chromatin ChIP for identifying ncRNA

+

- identification of regions that are transcribed
- reference-independent (if combined with ChIP-seq)
- detection of unstable transcripts if they have same chromatin state

-

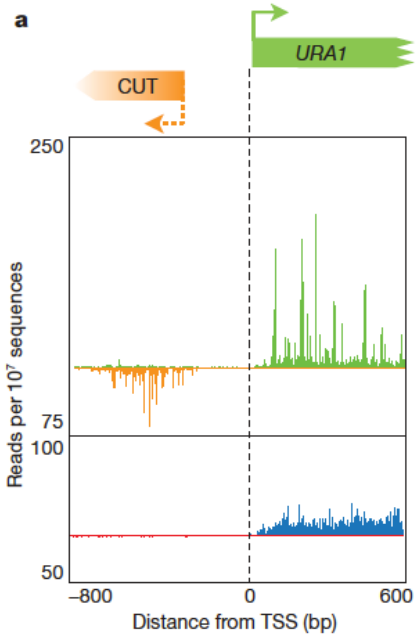
- indirect measure of expression

a) III: identification of unstable transcripts

- GRO-seq
- exosome knockout/ knock-down

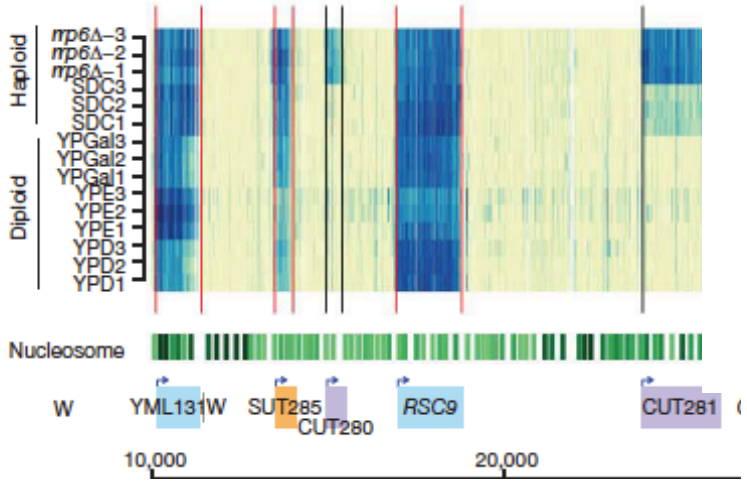
Nascent transcript sequencing visualizes transcription at nucleotide resolution

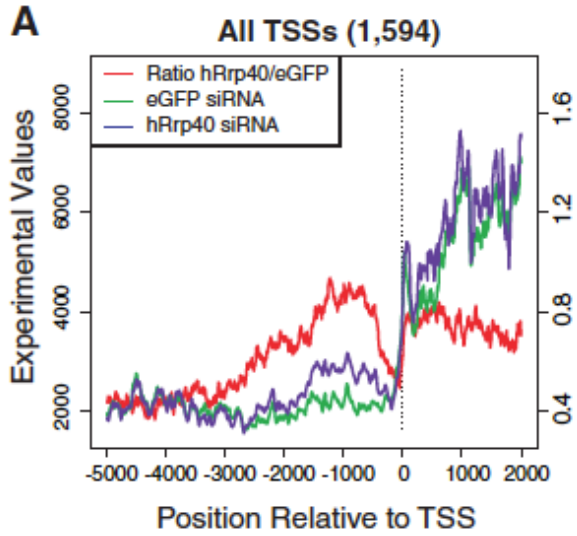
L. Stirling Churchman[1] & Jonathan S. Weissman[1]



RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters

Pascal Preker,[1] Jesper Nielsen,[2] Susanne Kammler,[1]* Søren Lykke-Andersen,[1] Marianne S. Christensen,[1] Christop





Bidirectional promoters generate pervasive transcription in yeast

Zhenyu Xu[1]*, Wu Wei[1]*, Julien Gagneur[1], Fabiana Perocchi[1], Sandra Clauder-Münster[1], Jurgi Camblong[2], Elisa Guffanti[3], Françoise Stutz[3], Wolfgang Huber[4] & Lars M. Steinmetz[1]

identification of unstable transcripts

+/- points for the individual techniques apply in addition

+

- comprehensive identification of transcripts

-

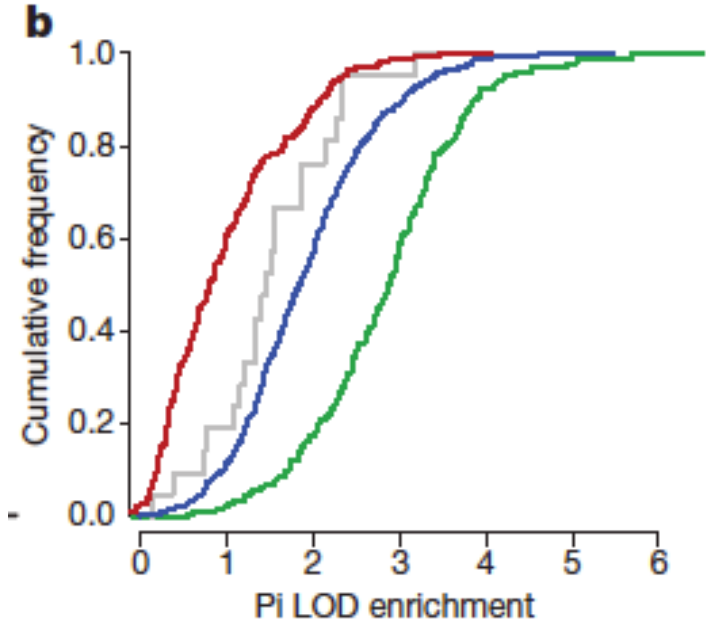
- challenge to assign functional significance

b) validation of non-coding transcripts

I. Are they functional
II. Are they non-coding

b) I. Validation: is the ncRNA functional

conservation: ncRNA are more conserved than random regions



red: protein coding
blue: lncRNA
green: random regions

Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals

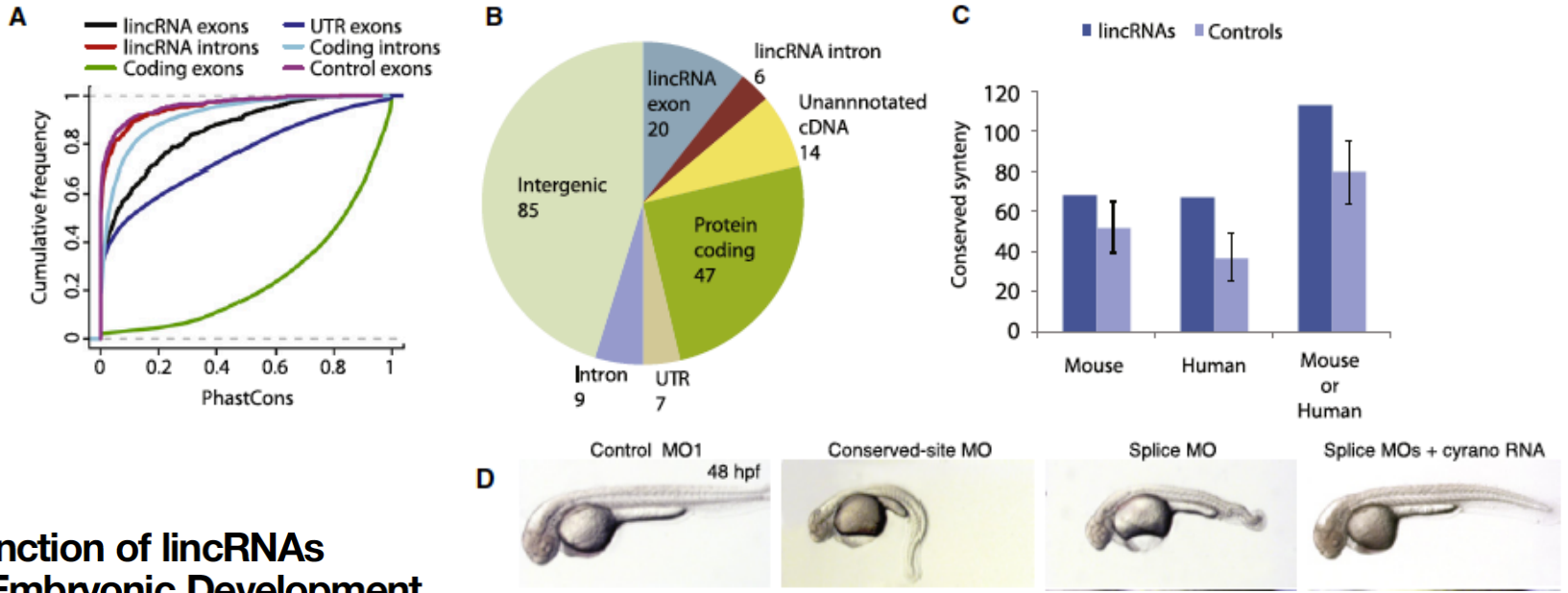
Mitchell Guttman[1,2], Ido Amit[1], Manuel Garber[1], Courtney French[1], Michael F. Lin[1], David Feldser[3], Maite Huarte[1,6], Or Zuk[1], Bryce W. Carey[7,8], John P. Cassady[2,8], Moran N. Cabili[7], Rudolf Jaenisch[2,8], Tarjei S. Mikkelsen[1,4], Tyler Jacks[2,3], Nir Hacohen[1,9], Bradley E. Bernstein[1,10,11], Manolis Kellis[1,5], Aviv Regev[1,2], John L. Rinn[1,6,11]* & Eric S. Lander[1,2,7,8]*

b) I. Validation: is the ncRNA functional - caveat

ncRNA more conserved than random regions

BUT location more conserved than sequence:

– e.g. possible to rescue ncRNA knockout zebrafish with mouse ncRNA



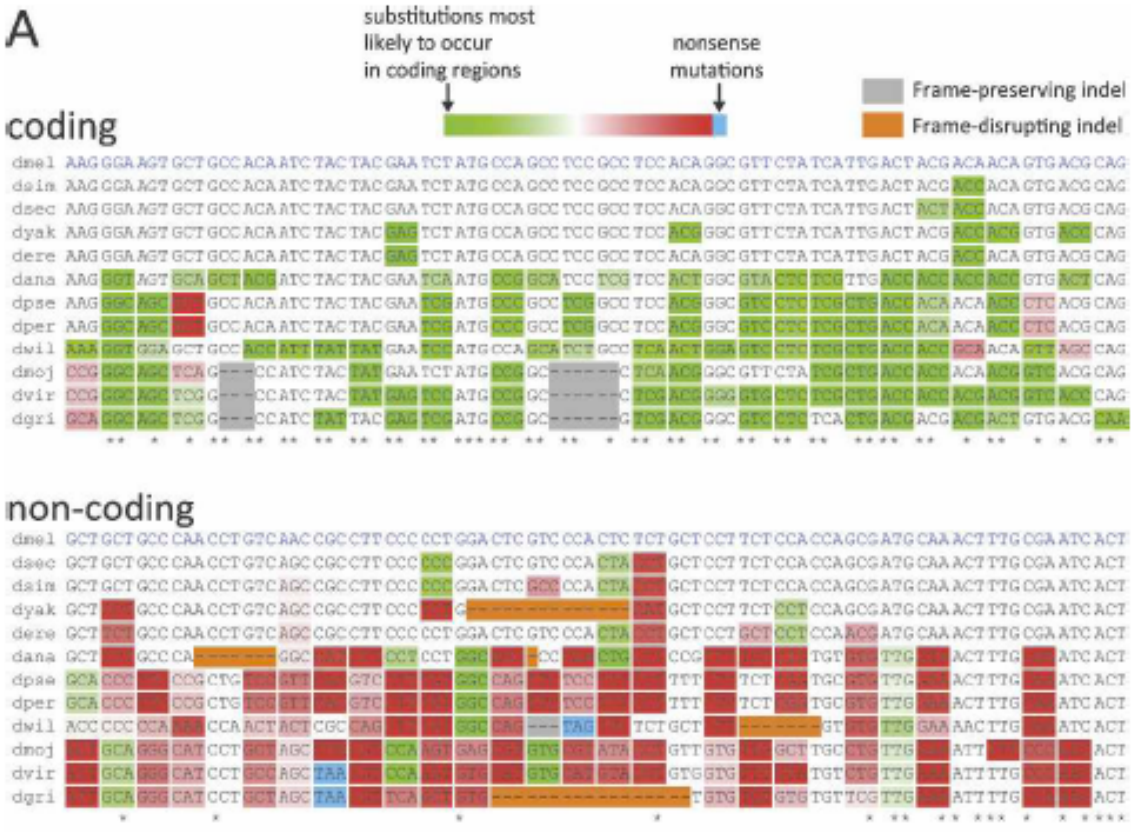
Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution

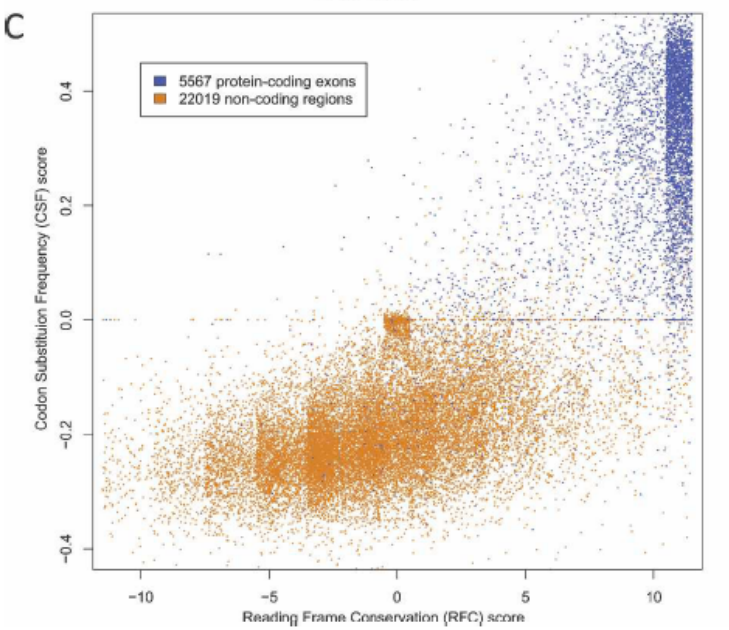
Igor Ulitsky,[1,2,3,5] Alena Shkumatava,[1,2,3,5] Calvin H. Jan,[1,2,3,4] Hazel Sive,[1,3] and David P. Bartel[1,2,3,*]

b) Validation: II. is the transcript non-coding?

- BLASTP the cDNAs sequences
- Assessment of codon substitution frequency
- Assessment of reading frame conservation

codon substitution frequency and reading frame conservation





12 *Drosophila* Genomes/Letter

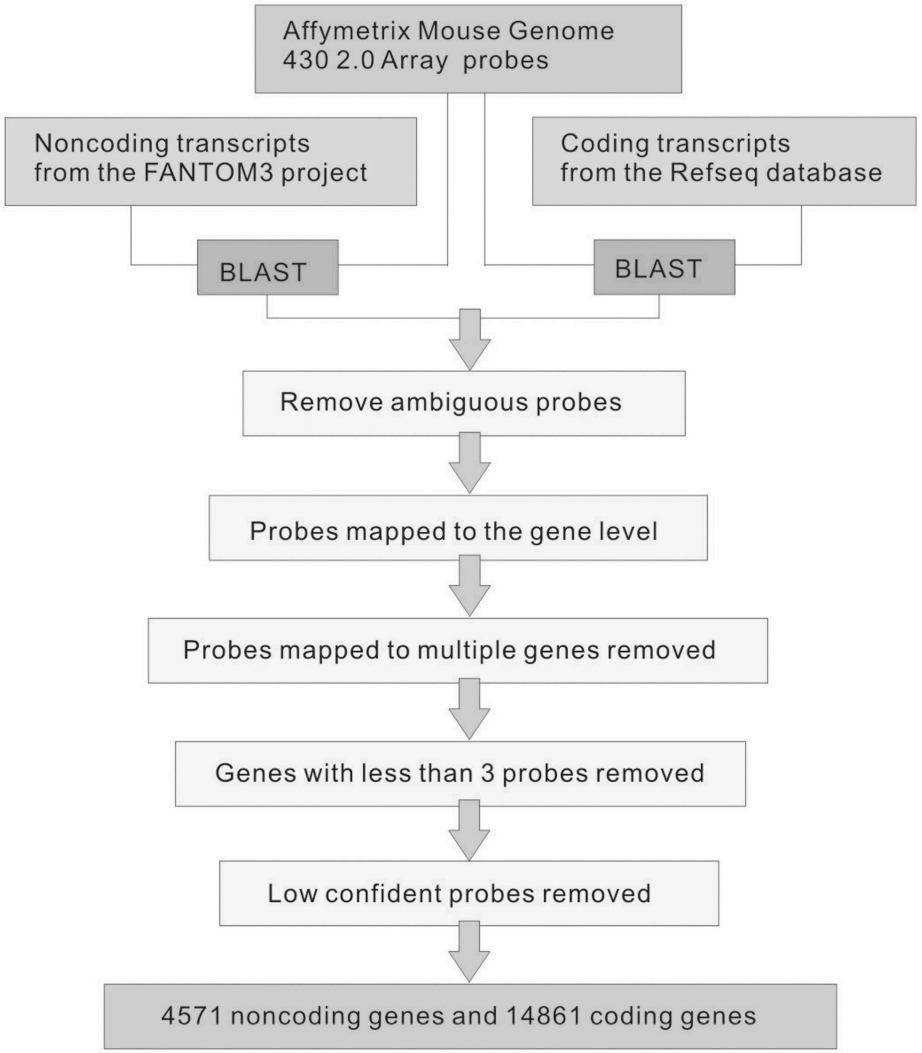
Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes

Michael F. Lin,[1] Joseph W. Carlson,[2] Madeline A. Crosby,[3] Beverley B. Matthews,[3] Charles Yu,[2] Soo Park,[2] Kenneth H. Wan,[2] Andrew J. Schroeder,[3] L. Sian Gramates,[3] Susan E. St. Pierre,[3] Margaret Roark,[3] Kenneth L. Wiley Jr.,[4] Rob J. Kulathinal,[3] Peili Zhang,[3] Kyl V. Myrick,[4] Jerry V. Antone,[4] Susan E. Celniker,[2] William M. Gelbart,[3,4] and Manolis Kellis[1,5,6]

- Within coding regions, triplet substitutions are biased toward conservative codon substitutions (Codon Substitution Frequencies, CSF).
- Indels in coding regions are strongly biased to be a multiple of three in length (reading frame conservation; RFC).

c) Computational identification based on published data: using microarray data – remapping the probes and literature curation

A



BIOINFORMATICS

Genome-wide computational identification and manual annotation of human long noncoding RNA genes

HUI JIA,[1] MAUREEN OSAK,[2] GIREESH K. BOGU,[3] LAWRENCE W. STANTON,[3] RORY JOHNSON,[3,4] and LEONARD LIPOVICH[1]

[1]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan 48202, USA
[2]Lee and Roland Witte Natural Sciences Division, Hillsdale College, Hillsdale, Michigan 49242, USA
[3]Stem Cell and Developmental Biology Group, Genome Institute of Singapore, 138672 Singapore

D210–D215 Nucleic Acids Research, 2012, Vol. 40, Database issue *Published online 1 December 2011*
doi:10.1093/nar/gkr1175

NONCODE v3.0: integrative annotation of long noncoding RNAs

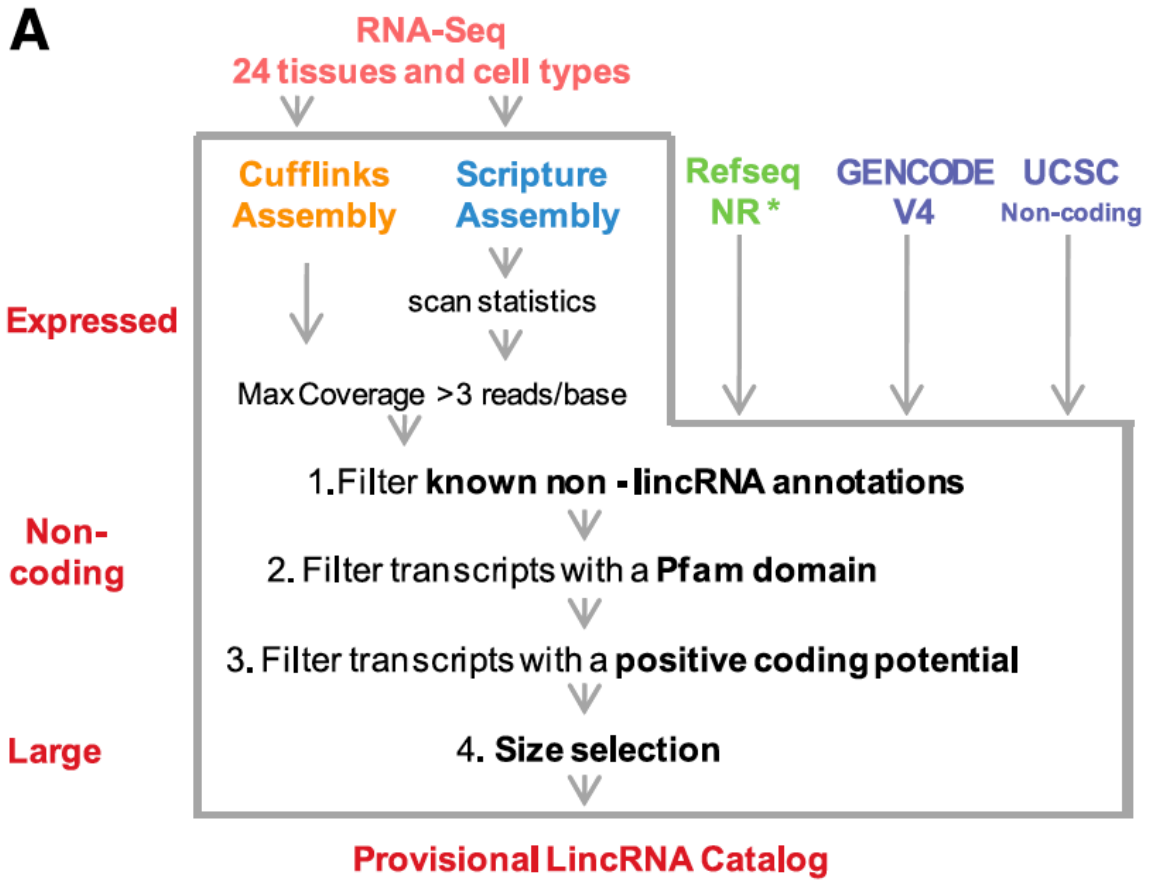
Dechao Bu[1,2], Kuntao Yu[1,2], Silong Sun[1], Chaoyong Xie[1,2], Geir Skogerbø[3], Ruoyu Miao[1,4], Hui Xiao[1], Qi Liao[1], Haitao Luo[1], Guoguang Zhao[1,2], Haitao Zhao[4], Zhiyong Liu[1], Changning Liu[1], Runsheng Chen[3,*] and Yi Zhao[1,*]

W118–W124 Nucleic Acids Research, 2011, Vol. 39, Web Server issue
doi:10.1093/nar/gkr432

ncFANs: a web server for functional annotation of long non-coding RNAs

Qi Liao[1,2,3], Hui Xiao[1], Dechao Bu[1,4], Chaoyong Xie[1], Ruoyu Miao[5], Haitao Luo[1], Guoguang Zhao[1,4], Kuntao Yu[1,4], Haitao Zhao[5], Geir Skogerbø[6], Runsheng Chen[6], Zhongdao Wu[2,3], Changning Liu[1,*] and Yi Zhao[1,*]

c) Computational identification based on published data: using RNA-seq data



Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses

Moran N. Cabili,[1,2,3] Cole Trapnell,[1,3] Loyal Goff,[1,4] Magdalena Koziol,[1,3] Barbara Tazon-Vega,[1,3] Aviv Regev,[1,5,6] and John L. Rinn[1,3,6,7]

Summary ncRNA identification

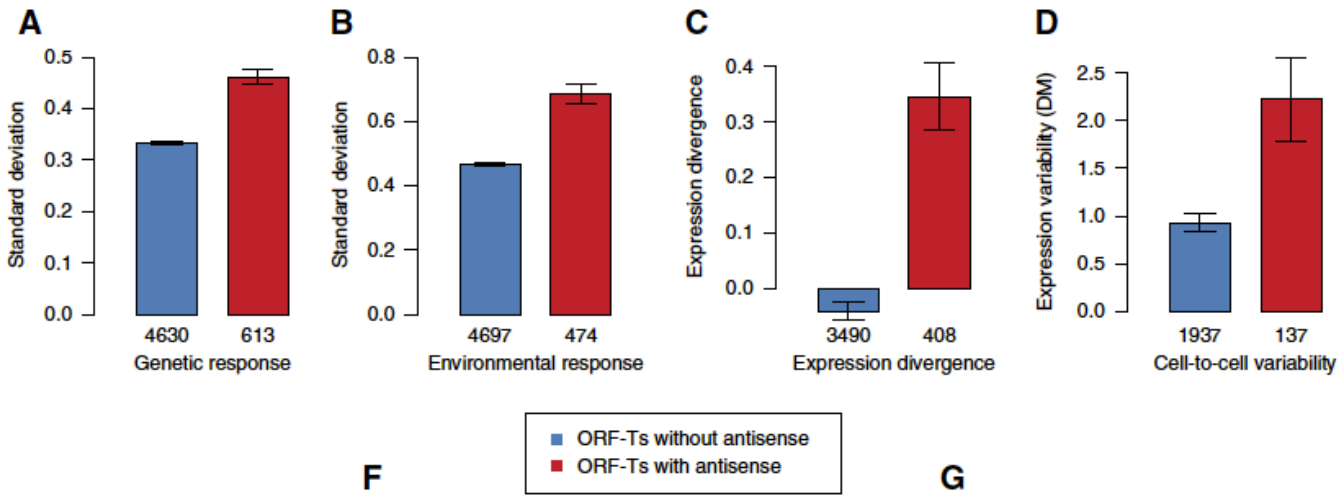
- Different template-free methods exist for detecting non-coding RNAs
- Methods to validate ncRNAs include
  - conservation: less than proteins, more than random – use positional conservation
  - codon substitution frequency
  - reading frame conservation
- Applications: meta studies integrating various published data sets to identify novel ncRNAs

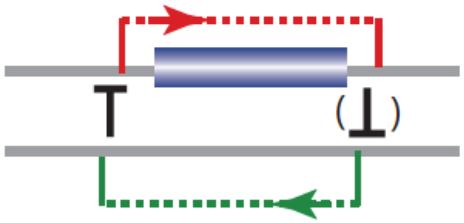
FUNCTIONAL IMPACT

Inferring the functional role of ncRNAs

a) Molecular mechanisms of ncRNAs activity
    I. sense antisense regulation
    II. association with protein complexes
    III. perturbation studies
b) Biological function
    I. tissue specificity
    II. functional analysis of genes co-expressed with ncRNAs

a) I. Sense antisense regulation: yeast





Molecular Systems Biology 7; Article number 468; doi:10.1038/msb.2011.1
Citation: Molecular Systems Biology 7:468

www.molecularsystemsbiology.com

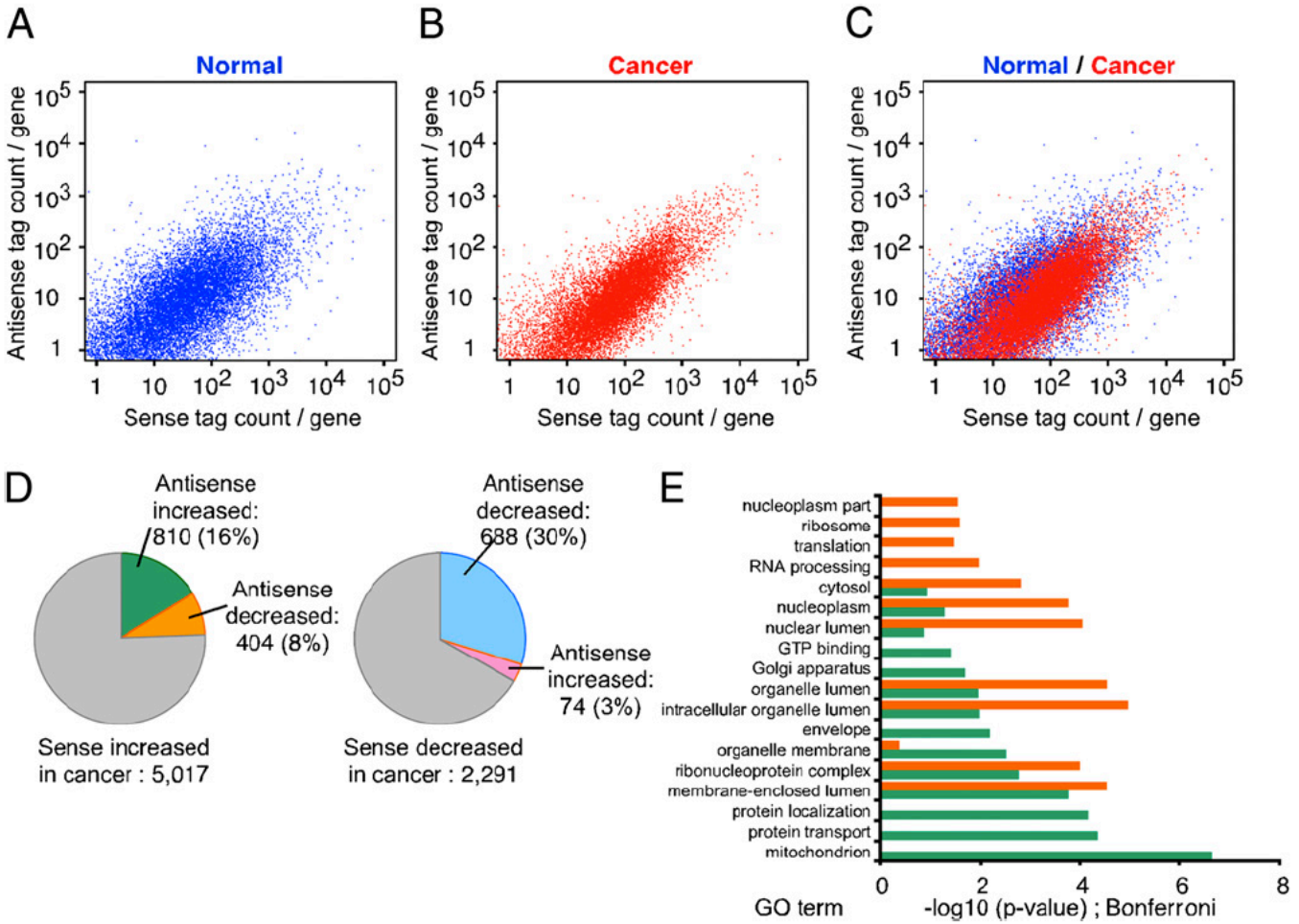
molecular systems biology

Antisense expression increases gene expression variability and locus interdependency

Zhenyu Xu[1], Wu Wei[1], Julien Gagneur[1], Sandra Clauder-Münster, Miłosz Smolik, Wolfgang Huber and Lars M Steinmetz*

Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[1] These authors contributed equally to this work
* Corresponding author. Genome Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany.
Tel.: +49 6221 387 8389; Fax: +49 6221 387 8518; E-mail: larsms@embl.de

a) I. Sense antisense regulation: breast cancer

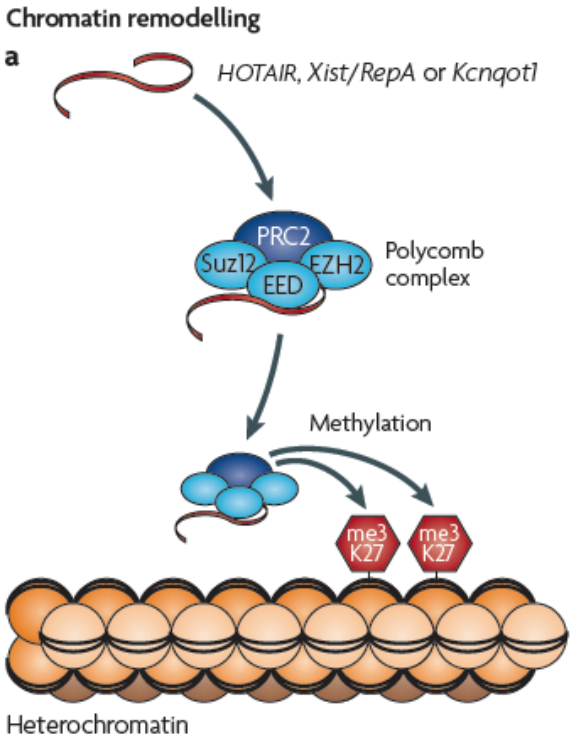


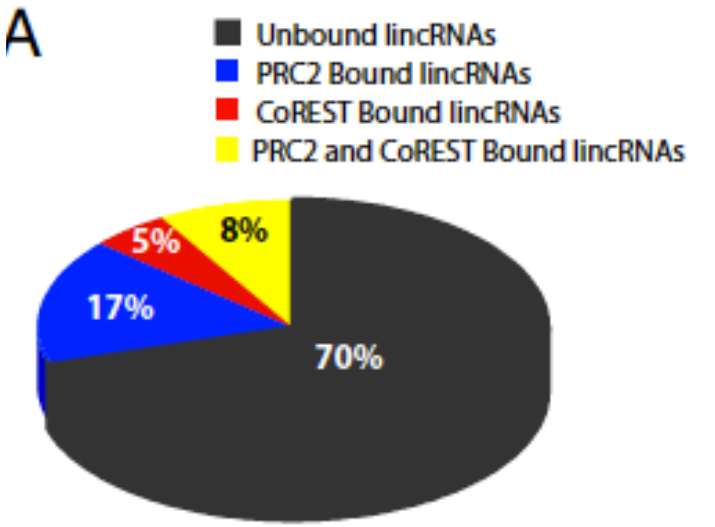
Altered antisense-to-sense transcript ratios in breast cancer

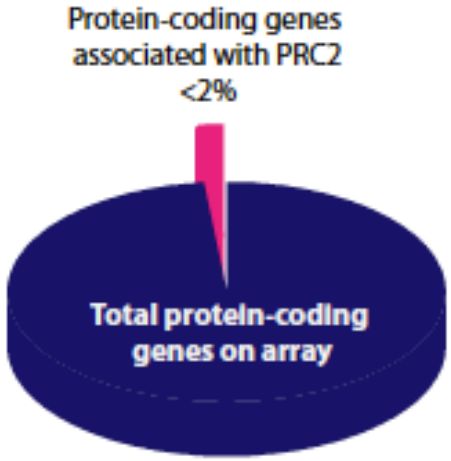
Reo Maruyama[a,b,c,1], Michail Shipitsin[a,b,c,1], Sibgat Choudhury[a,b,c,1], Zhenhua Wu[d,e,1], Alexei Protopopov[f], Jun Yao[g], Pang-Kuo Lo[h], Marina Bessarabova[i], Alex Ishkin[i], Yuri Nikolsky[j], X. Shirley Liu[d,e], Saraswati Sukumar[h], and Kornelia Polyak[a,b,c,k,2]

a) II. association with protein complexes

Inspired by example







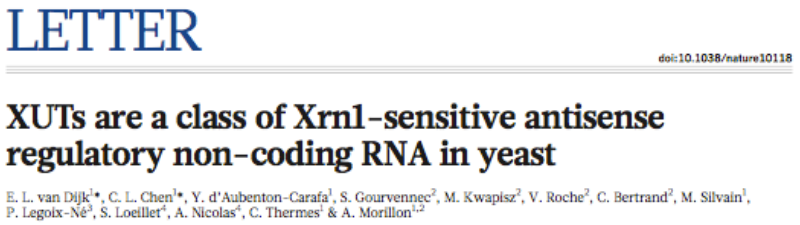
Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression

Ahmad M. Khalil[a,b,1], Mitchell Guttman[a,c,1], Maite Huarte[a,b], Manuel Garber[a], Arjun Raj[d], Dianali Rivea Morales[a,b], Kelly Thomas[a,b], Aviva Presser[a], Bradley E. Bernstein[a,e], Alexander van Oudenaarden[d], Aviv Regev[a,c], Eric S. Lander[a,c,f,1,2], and John L. Rinn[a,b,1,2]

The Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142; [b]Department of Pathology, Beth Israel Deaconess

a) III. perturbation studies: special case yeast

- ncRNAs exported to cytoplasm



LETTER

doi:10.1038/nature10118

XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast

E. L. van Dijk[1]*, C. L. Chen[1]*, Y. d'Aubenton-Carafa[1], S. Gourvennec[2], M. Kwapisz[2], V. Roche[2], C. Bertrand[2], M. Silvain[1], P. Legoix-Né[3], S. Loeillet[4], A. Nicolas[4], C. Thermes[1] & A. Morillon[1,2]

- ncRNAs specifically stabilized in meiosis required

Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6

Aurélie Lardenois[a,1], Yuchen Liu[a,1], Thomas Walther[b], Frédéric Chalmel[a], Bertrand Evrard[a], Marina Granovskaia[c], Angela Chu[d], Ronald W. Davis[d,e,2], Lars M. Steinmetz[c], and Michael Primig[a,2]

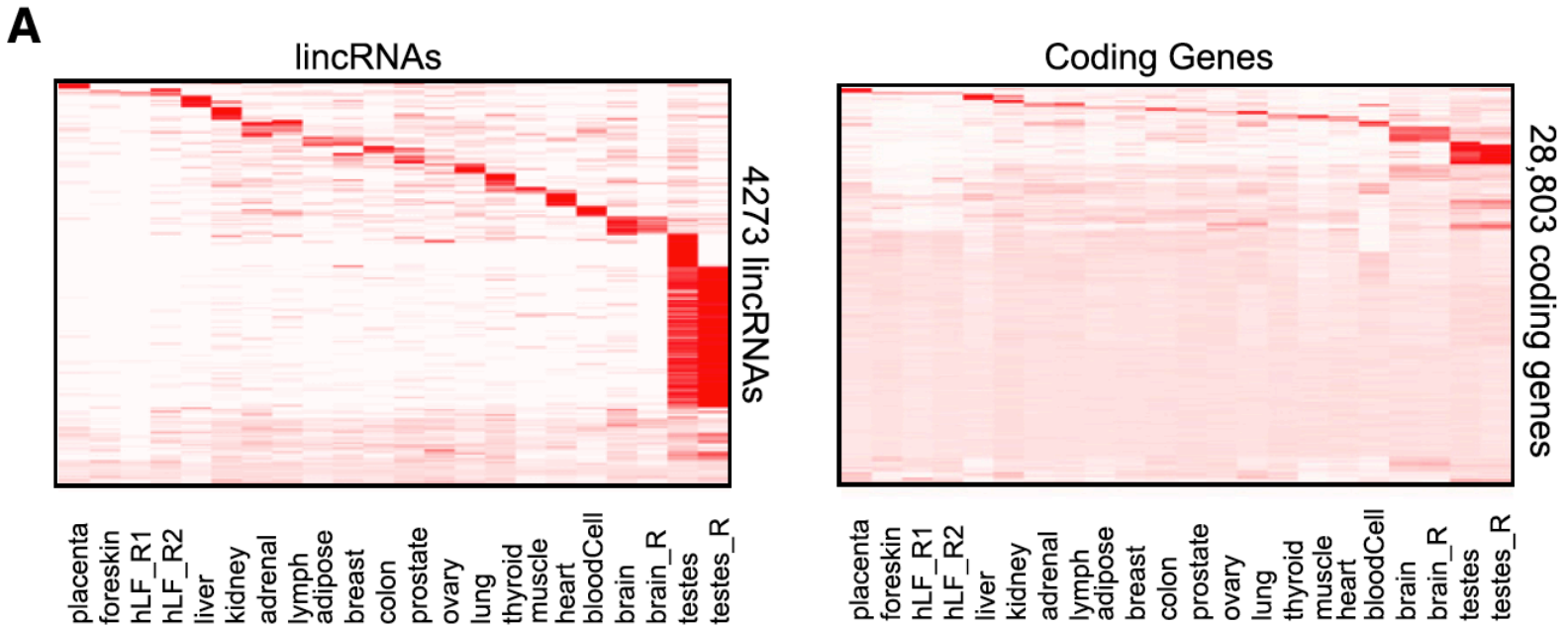
- ncRNAs only detectable in exosome knockouts

LETTERS

Bidirectional promoters generate pervasive transcription in yeast

Zhenyu Xu[1]*, Wu Wei[1]*, Julien Gagneur[1], Fabiana Perocchi[1], Sandra Clauder-Münster[1], Jurgi Camblong[2], Elisa Guffanti[3], Françoise Stutz[3], Wolfgang Huber[4] & Lars M. Steinmetz[1]

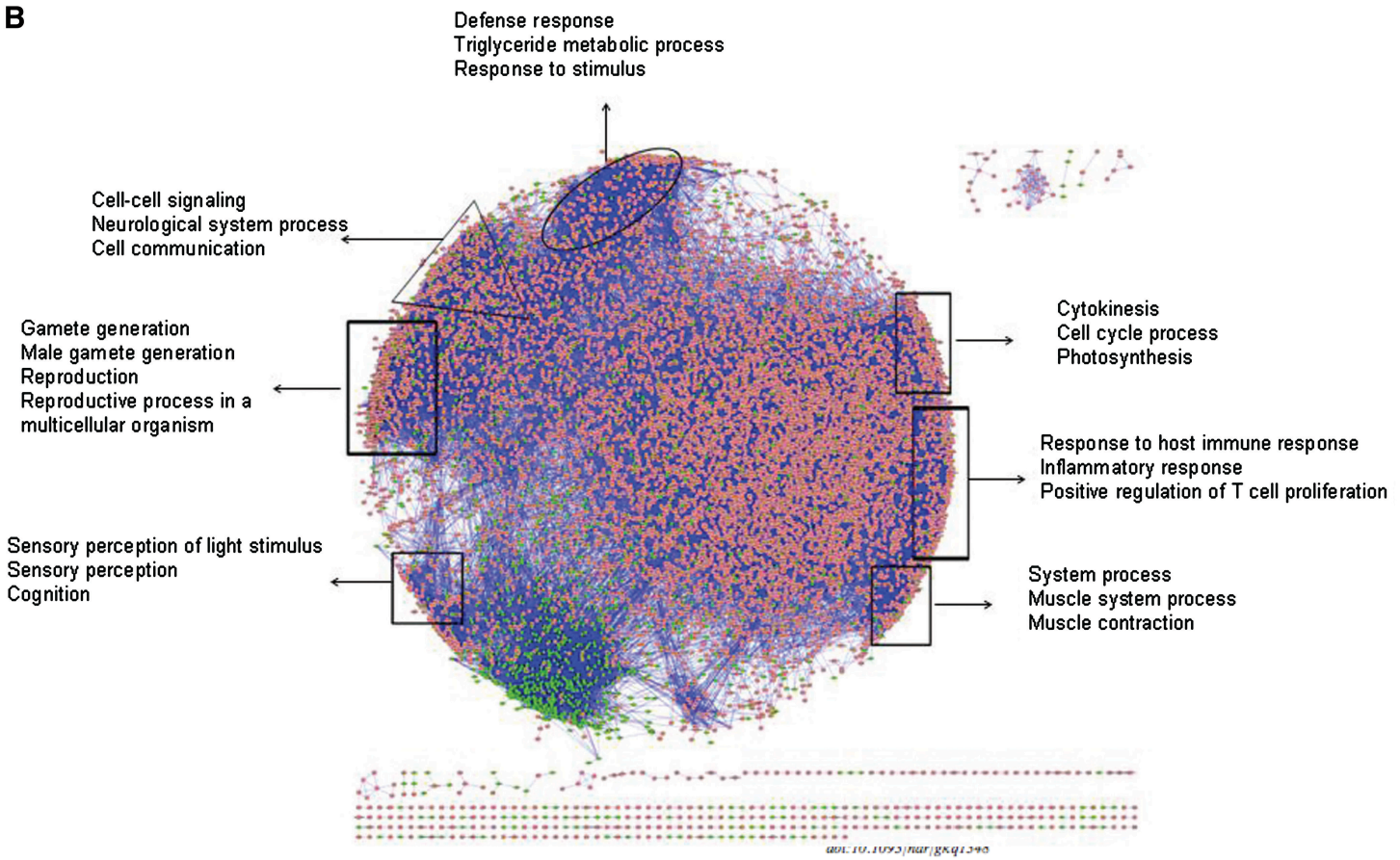
b) I. Tissue-specificity



Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses

Moran N. Cabili,[1,2,3] Cole Trapnell,[1,3] Loyal Goff,[1,4] Magdalena Koziol,[1,3] Barbara Tazon-Vega,[1,3] Aviv Regev,[1,5,6] and John L. Rinn[1,3,6,7]

b) II. functional analysis of genes co-expressed with ncRNAs: pairwise correlation with genes

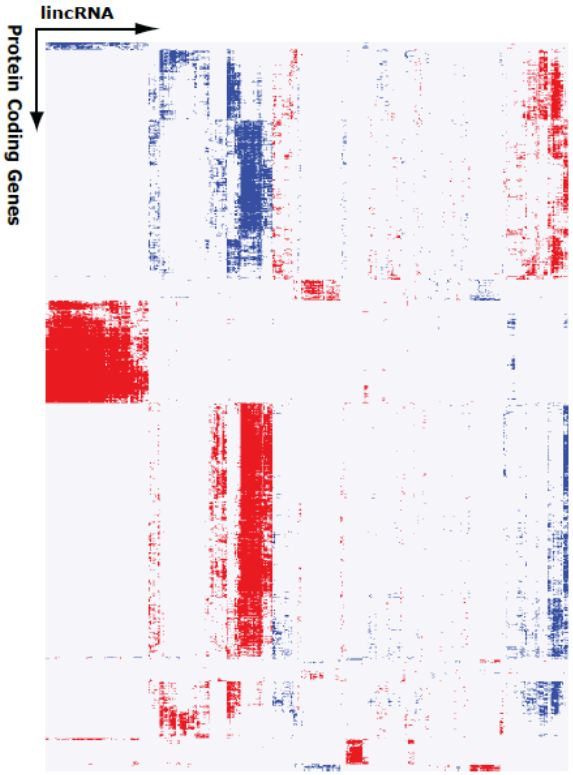


Published online 18 January 2011

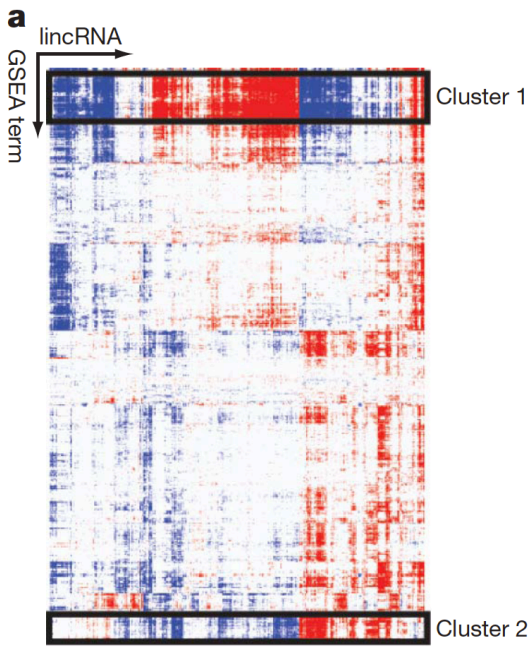
Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network

Qi Liao[1,2,3], Changning Liu[1], Xiongying Yuan[1,4], Shuli Kang[1], Ruoyu Miao[5], Hui Xiao[1], Guoguang Zhao[1,4], Haitao Luo[1], Dechao Bu[1,4], Haitao Zhao[5], Geir Skogerbø[6], Zhongdao Wu[2,3,*] and Yi Zhao[1,*]

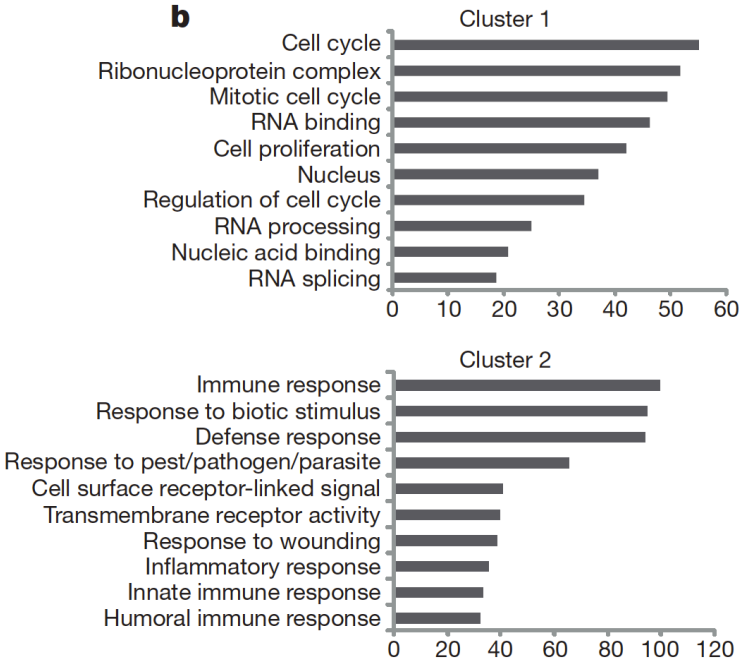
Correlation of ncRNA and proteins across 19 conditions



Correlation ncRNA with gene sets (sets of proteins) and clustering



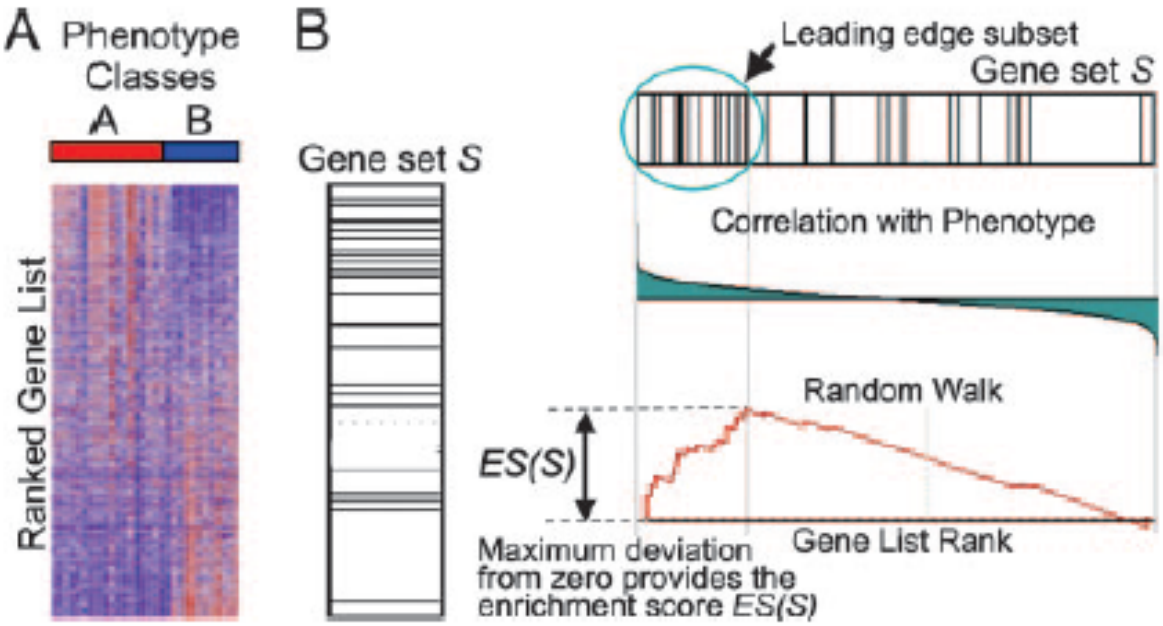
GSEA reveals biological function of ncRNAs



Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals

Mitchell Guttman[1,2], Ido Amit[1], Manuel Garber[1], Courtney French[1], Michael F. Lin[1], David Feldser[3], Maite Huarte[1,6], Or Zuk[1], Bryce W. Carey[2,8], John P. Cassady[2,8], Moran N. Cabili[2], Rudolf Jaenisch[2,8], Tarjei S. Mikkelsen[1,4], Tyler Jacks[2,3], Nir Hacohen[1,9], Bradley E. Bernstein[1,10,11], Manolis Kellis[1,5], Aviv Regev[1,2], John L. Rinn[1,6,11]* & Eric S. Lander[1,2,7,8]*

Excursion: Gene set enrichment analysis



Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

Summary of functional annotation

- strategy for uncovering mechanisms: generalization of examples
  - diverse molecular mechanism: antisense, association with complexes
- strategy for uncovering biological function: association with known proteins
  - ncRNAs are expressed very tissue-specifically

Conclusions

- **Many different technologies exist for identifying ncRNA**
  - based on the method you find different ncRNA types
  - be aware when comparing different data sets!
- **Computational data mining of published data has been successfully applied to identify novel ncRNAs**
  - generally identify more transcripts than single studies
  - less sequence conservation than proteins makes it more challenging to identify functional ncRNAs (use positional conservation)
  - be aware that many ncRNAs are tissue-specific!
- **Functional insights are starting to emerge**
  - generalizing known examples has been successful
  - associations of ncRNAs and genes can give insights into biological functions
  - nevertheless, we are still far away from understanding them as we do proteins